Journal of Epidemiology and Community Health, 1986, **41**, 55–58

Epidemiology of AIDS—statistical analyses

OLE-JAN IVERSEN[1] AND STEINAR ENGEN[2]

From the Department of Microbiology[1] and the Department of Mathematics and Statistics,[2] University of Trondheim, Norway

SUMMARY Some central questions concerning the epidemiology of AIDS are addressed by statistical analyses. Applying standard maximum likelihood theory to reported cases of transfusion-associated AIDS in the US, the mean and standard deviation of incubation time for AIDS are estimated to be about 60 and 19 months, respectively. If these parameters are applied to the data from the San Francisco CDC cohort study, we find a good correspondence between estimated and reported cases of AIDS when the probability factor p is 0·27—meaning that about 27% of those infected with HIV are expected to develop AIDS during a period of 8–10 years. Application of the incubation time model and the probability factor p to the data on transfusion-associated AIDS makes it possible to estimate the number of transfusion-associated infections with HIV from 1978 to 1984. These estimates give an exponential increase in the number of cases, with a relative increase of 2·74 each year. It seems reasonable to assume that this increase reflects the spread of the virus within this period.

The aetiological agent of AIDS is identified as a retrovirus designated human immunodeficiency virus (HIV). Considerable knowledge about the virus and how it acts has emerged during the last few years.[1] This does not, however, mean that we are close to a solution of the AIDS problem. The research has revealed a high degree of genomic diversity between different HIV isolates.[1] The most divergent part of the genomes lies in the *env* gene, resulting in great divergence in the amino acid composition of the envelope proteins.[1] It is therefore unlikely that an effective vaccine against HIV will be available in the near future.

Meanwhile the AIDS epidemic goes on. During the first 20 weeks in 1986, 4762 cases of AIDS were registered at the Center of Disease Control (CDC) in the United States.[2] The corresponding figures for the same period in 1985 and in 1984 were 2655 and 1441, respectively.[2 3] Thus, the relative increase in cases of AIDS is nearly constant.

What about the spread of HIV infections? Owing to the considerable lag between the virus infection and the occurrence of AIDS, the epidemiological data do not reflect the spread of the virus at the current time. What then is the lag between HIV infection and the occurrence of AIDS, and how many of those infected are expected to develop AIDS? In this communication answers to these central questions in the epidemiology of AIDS are sought through statistical analyses of published data.

Estimates of the lag between HIV infection and the occurrence of AIDS

In order to estimate the incubation time for AIDS, the point in time of the infection must be known. This is ascertainable for transfusion-associated AIDS. Within this group the point of time for infection as well as for AIDS diagnosis is known with sufficient precision. We have analysed the data set presented by Peterman *et al*[4] reporting about 150 cases of AIDS caused by blood transfusion during the period 1978–84.

It seems reasonable to assume that the time from infection to AIDS diagnosis, say T, is a random variable with a certain probability distribution. However, these data are not observations from this distribution but are necessarily censored. For example, AIDS diagnoses made before 1984 among those infected in 1980 can include only patients who developed the syndrome within a period of four years after infection. These observations are therefore generated by the conditional distribution of T given the event $T < 4$. Data from each year are observations from different conditional distributions of this type. A preliminary rough analysis of the distribution of T over intervals of lengths one year indicates that the unconditional distribution of T is a unimodal, symmetric distribution resembling the Gaussian curve. We therefore make the assumption that T is normally distributed with density

J Epidemiol Community Health: first published as 10.1136/jech.41.1.55 on 1 March 1987. Downloaded from http://jech.bmj.com/ on September 24, 2023 by guest. Protected by copyright.

$$f(t)=\frac{1}{\sqrt{2\pi}}\,\frac{1}{\sigma}\,e^{-\left(\frac{t-\mu}{\sigma}\right)^2}$$

In this expression $\mu=E(T)$ and $\sigma^2=\text{var}(T)$ are the mean and the variance, respectively, of the incubation time T.

In fact it is sufficient for the estimation of μ and σ to assume that cases with $(T<8)$ follow this distribution truncated at $T=8$ years. Since there are at present no observed cases with incubation times of more than eight years, no information about the distribution to the right of eight years can be drawn from the sample. However, the data indicate that a certain fraction of those infected develop AIDS according to some model which approximates the normal distribution. It remains to be seen what the right tail of the distribution really is. Those who have not developed AIDS within eight years may not develop AIDS at all, or may possibly develop it some time much later than eight years.

Mathematically, the unknown right tail of the distribution may be included by writing

$$p\,f(t)+(1-p)w(t)$$

for the density of T, where p is the proportion following the normal distribution f(t), and w(t) is the distribution of incubation times greater than eight years that will remain unknown. If only those following the normal model develop AIDS, w(t) will have its probability mass concentrated at infinity. In our estimation procedure, including only data with $T<8$, only the distribution f(t) will occur.

No simple estimates for μ and σ^2 exist when the data are generated as described above. However, using numerical mathematical techniques, the maximum likelihood estimators may be found, though the computing time is quite long. The computations are carried out on a VAX machine at AVH, the University of Trondheim, and give estimates which are approximately

$\mu=60$ (months)
$\sigma=19$ (months)

The estimates seem surprisingly stable in time. For example, estimates for μ and σ computed on the bases of data collected before 1984 (58 cases only) are very close to those found from the total set of data.

Estimates of the proportion of those infected with HIV who develop AIDS

The data set presented by Peterman *et al*[4] concerning transfusion associated AIDS does not say anything about the number of people who are actually infected with HIV by blood transfusion. The random variable T refers to those among the infected who develop AIDS. To estimate the proportion of those infected with HIV who will develop AIDS within a period of 8–10 years we have combined the incubation time model of the previous section with the San Francisco CDC cohort study presented by Curran *et al*..[5] In this study, we know the estimated number of infected persons and the registered cases of AIDS (table 1).

Let g(t) denote the number of those infected within this group as a function of time. Then, according to our model, the expected number of AIDS cases is

$$h(t)=\int_{-\infty}^{t} p\,g'(u)\,\Phi\left(\frac{t-u-\mu}{\sigma}\right)du$$

$$+\int_{-\infty}^{t-96}(1-p)\,g'(u)W(t-u)du$$

here p is the proportion who develop AIDS according to the normal model and $\Phi(\cdot)$ is the standard normal integral

$$\Phi(x)=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}u^2}du$$

and

$$W(x)=\int_{-\infty}^{x} w(u)du.$$

Since $g'(u)$ is very small for small values of u, the last integral in the expression for h(t) will give only a very small contribution and may be neglected. In fact, $g'(u)$ for u-values contributing to this integral is the absolute growth rate for the number of those infected more than 8 years ago.

For the San Francisco cohort, g(t) is very close to a straight line over several years, that is, $g'(t)$ is a

Table 1 *Estimated and reported cases of AIDS in the San Francisco CDC cohort study*

Year	*1978*	*1979*	*1980*	*1981*	*1982*	*1983*	*1984*
Estimated number of HIV seropositive	275	825	1650	2406	3162	3919	4675
Cumulative number reported with AIDS	0	0	2	14	41	84	166
Estimated AIDS cases (cum) $\mu=60$, $=19$:							
$p=0{\cdot}25$	0	0·3	2·1	9·5	31·1	77·6	154·1
$p=0{\cdot}27$	0	0·3	2·3	10·3	33·6	83·8	166·4
$p=0{\cdot}30$	0	0·4	2·6	11·4	37·4	93·1	185·0

J Epidemiol Community Health: first published as 10.1136/jech.41.1.55 on 1 March 1987. Downloaded from http://jech.bmj.com/ on September 24, 2023 by guest. Protected by copyright.

constant. Inserting this constant and our estimates μ^* and σ^*, we can study the fit to the registered cases of AIDS (table 1). The deviation between registered and estimated values is minimised for $p = 0{\cdot}27$. Altogether, our estimates indicate that between 25 and 30% of those infected with HIV are expected to develop AIDS within 8–10 years.

The statistical uncertainty in μ^* will affect the uncertainty in p. If we insert μ-values in the interval (55,65), the corresponding estimates of p are in the interval (0·22,0·35).

Estimates of the number of persons infected with HIV by blood transfusion

We may now estimate the number of persons infected by blood transfusion in the US. Let N_t denote the number infected by blood transfusion during a time interval around time t (year t). Among this group, let the number of AIDS cases reported before time s be denoted $Z_{s,t}$. Then, according to the model

$$EZ_{s,t} = N_t \cdot q_{s,t}$$

where

$$q_{s,t} = p\Phi\left(\frac{s-t-\mu}{\sigma}\right)$$

inserting our estimates for p,μ and σ we obtain an estimate $\hat{q}_{s,t}$ for $q_{s,t}$. Then, N_t can be estimated as

$$\hat{N}_t = Z_{s,t} \cdot \hat{q}_{s,t}^{-1}$$

Conditional on N_t, the variable $Z_{s,t}$ is binomially distributed with parameters $(Nt, q_{s,t})$.

Hence

$$\text{var}(Z_{s,t}) = N_t\, q_{s,t}\,(1 - q_{s,t})$$

and consequently

$$\text{var}(\hat{N}_t) = N_t(1 - q_{s,t})/q_{s,t}.$$

Using first order Taylor approximation we then find

$$\text{var}(1n\hat{N}_t) = (1 - q_{s,t})/(q_{s,t}N_t)$$

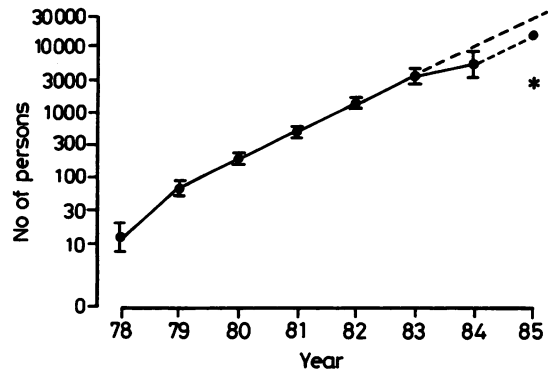
which can be estimated by $(1 - \hat{q}_{s,t})/Z_{s,t}$.

The estimates of N_t plotted on a logarithmic scale against t are given in the figure. The estimated values are close to a straight line, corresponding to exponential growth.

The best linear estimate of the parameters based on the data from 1978 to 1983, taking into account that the $1n(N_t)$ have different variances, gives the relation

$$\hat{N}_t = 23{\cdot}15e^{1{\cdot}01(t-1978)}$$

where t is the time measured in years. The χ^2-statistic for testing goodness of fit turns out to be 2·16 with 4 degrees of freedom, showing a very good fit to the model.



Estimates of the number of transfusion-associated infections with HIV in the US. The asterisk (*) indicates the expected number of cases in 1985 in spite of serological screening (see text).

Notice that the estimated number of persons infected by blood transfusion increases by a nearly constant factor

$$\hat{k} = e^{1{\cdot}01} = 2{\cdot}74$$

from one year to the next. The 95% confidence interval for k is [2·44, 3·09].

The estimate of infected persons in 1983 is 3600, and with a relative increase of 2·74 the expected number for 1984 would be 9900. However, based on the registered cases of AIDS, our estimates suggest only 5700 infected persons. This moderate increase is most likely a result of recommendations from the Public Health Service that members of groups at high risk for AIDS should refrain from donating blood or plasma.

Discussion

Some central questions concerning the epidemiology of AIDS are addressed by statistical analyses. We have used the data from transfusion-associated cases of AIDS in the US[4] to estimate the incubation period for AIDS. For this group we are able to identify the time of exposure for HIV as well as the time of diagnosis. Applying standard maximum likelihood theory to these data, the mean and standard deviation of the incubation time for AIDS are estimated to be about 60 and 19 months, respectively.

J Epidemiol Community Health: first published as 10.1136/jech.41.1.55 on 1 March 1987. Downloaded from http://jech.bmj.com/ on September 24, 2023 by guest. Protected by copyright.

Our next question was what proportions (p) of those infected with HIV will develop AIDS. The San Francisco CDC cohort study gives us data for seroconversion (partly estimated) and reported cases of AIDS over a period of seven years. When applying our incubation time model to these data, we find a very good correspondence between estimated and reported cases of AIDS if the probability factor p is 0·27, meaning that 27% of those infected are expected to develop AIDS during a period of 8 to 10 years.

With our incubation time model and our estimates of the probability factor we are now able to estimate the number of recipients of HIV infected blood in the US in the period 1978 to 1984. Our data presented in the figure show an exponential increase in the number of infected persons from 1978 to 1983. The number of HIV recipients increases by a factor of 2·74 each year. As far as we can understand this must reflect the spread of the virus in the population and an increase in the proportion of infected donors. Any increase in the number of persons receiving blood during this period is probably small and will have little influence on these data. If the number of persons infected with HIV increases by a factor of 2·74 each year, this means that the number of HIV infected persons doubles every 8·2 months.

Since early 1985 blood donors have been tested for antibodies to HIV. This has certainly reduced the risk for transfusion-associated AIDS. However, among those infected with HIV some will be seronegative, because of either early infection with the virus or the presence of high titres of virus.[6] The period from infection to seroconversion varies considerably. Let us suppose a mean incubation time for seroconversion of two months. The proportion of those infected for less than x months can be estimated from

$$1 - e^{-1 \cdot 01 \cdot x/12}$$

If x is 2, it gives 0·155, meaning that 15·5% have been infected less than two months and are still seronegative. Let us now return to the figure. The number of persons infected by transfusion in 1984 is estimated at 5700. With a relative increase of 2·74 the expected number for 1985 would be about 15 600. It would follow that 15·5% or 2400 (figure) were infected in spite of serological screening because the infected donors were still seronegative. The asterisk (*) in the figure represents predicted infections for 1985 even if serological screening were complete and accurate. This emphasises how important it is that persons at increased risk of HIV infection should refrain from donating blood.

References

[1] CDS. Cases of specified notifiable diseases, United States. *MMWR* 1986; **35:** 209.

[2] CDS. Cases of specified notifiable diseases, United States. *MMWR* 1985; **34:** 176.

[3] Wong Staal F, Gallo RC. Human T-lymphotropic retroviruses. *Nature* 1985; 395–403.

[4] Peterman TA, Jaffe HW, Feorino PM, Getchell JP, Warfield DT, Haverkos HW, Stoneburner RL, *et al.* Transfusion-associated acquired immunodeficiency syndrome in the United States. *JAMA* 1985; **254:** 2913–7.

[5] Curran JW, Morgan WM, Hardy AM, Jaffe HW, Darrow WW, Dowöle WR. The epidemiology of AIDS: Current status and future prospects. *Science* 1985; **229:** 1352–7.

[6] Levy JA, Kaminsky LS, Morrow WJW, Steimer K, Luciw P, Dina D, Hoxie J, *et al.* Infection by the retrovirus associated with the acquired immunodeficiency syndrome. *Ann Intern Med* 1985; **103:** 694–9.

J Epidemiol Community Health: first published as 10.1136/jech.41.1.55 on 1 March 1987. Downloaded from http://jech.bmj.com/ on September 24, 2023 by guest. Protected by copyright.